Bibliográfiai hivatkozások automatikus kinyerése[1]

Váradi Tamás[1], Pintér Tibor[1], Mittelholcz Iván[1], Peredy Márta[1]

[1] MTA Nyelvtudományi Intézet
Benczúr utca 33., 1068 Budapest
{varadi, tpinter, mittelholcz, mperedy}@nytud.hu

Kivonat: A Magyarországon megjelentetett társadalomtudományi folyóiratok tanulmányaiból automatikusan kigyűjtött hivatkozások adatbázisba rendezése jelentős segítség a tudomány számára. A heterogén források által többféle struktúrában megjelenített adatok elemzését és azonos formátumba rendezését a szabad felhasználású NooJ szoftver segítségével végeztük. A folyamat valódi kihívása az adathalmaz elemeinek, valamint a hivatkozások típusának automatikus felismerésében rejlik. A külön-külön létrehozott (ugyanakkor egymással kombinálható) NooJ-grammatikák szerepe a hivatkozások egyes elemeinek felismerése és annotálása. Az automatizált folyamat kimeneteként létrejövő XML-elemek még utólagos kézimunkára szorulnak, részint a hivatkozások rossz minősége miatt (hiányos hivatkozások, szabványoktól eltérő hivatkozások), részint a folyamat formalizált volta miatt (bizonyos hivatkozások automatikusan több hivatkozástípusba is besorolódnak). A BibTex-szabványosítás előtt egyértelműsítő algoritmusokat és/vagy kézi erőt kell használni.

1 Háttér

Munkánk célja a magyar társadalomtudományi folyóiratokban történő hivatkozások adatbázisának létrehozása és folyamatos, automatizált bővítése. Az MTA Nyelvtudományi Intézetben az OTKA támogatásával végzett munka jelentőségét az adja, hogy Magyarországon nem létezik ilyen átfogó adatbázis (bár kétségtelen, hogy hasonlók léteznek, ám lefedettségben és a létrehozás módszerében projektünk ez idáig Magyarországon egyedülálló). Egy ilyen jellegű adatbázis nagyon hasznos szerepet tölthetne be a társadalomtudományi folyóiratok színvonalának, valamint a kutatók publikációs tevékenységének megbízható elbírálásában.

Az adatok mennyisége és sokfélesége nélkülözhetetlenné teszi a korszerű számítógépes technológia használatát. A magyarországi hivatkozás-adatbázisok építésének (MTMT adatbázis) gyakorlatában főként a kézzel történő adatbevitel van elterjedve, ami hosszútávon nem járható út, mivel időigényes és a manuális adatbevitel miatt nehézkes (nem beszélve a többletmunkáról, hiszen az adatbázisba egy már írásban megjelent adathalmazt írnak be – manuálisan).

[1] Jelen munka az OTKA PUB-F, 81666 számú pályázat keretében készült.

2 A feladat

A hivatkozások automatikus kigyűjtésének alapvetően két módja van: a mintaalapú szövegkinyerés (alapvetően előre létrehozott karaktersorok alapján történő szegmentálás és adatkinyerés) és a gépi tanulási technikákon, valamint statisztikai módszereken alapuló szövegbányászat [1, 2, 3, 4, 5]. Az OTKA által támogatott projektünkben elsősorban karakteralapú szövegfelismerést alkalmazzuk, melyet a hivatkozások állandó elemeiből összeállított szótárakkal egészítünk ki (egyúttal további szótárak létrehozását is tervezzük).

A munka az alábbi szakaszokra bontható:

i. a társadalomtudományi folyóiratok körének feltérképezése, a bennük található hivatkozások szöveges alakban történő összeállítása,
ii. a szabad szövegű hivatkozások bibliográfiai elemeinek (pl. szerző, cím, kiadó, kiadás helye stb.) automatikus azonosítása és annotálása,
iii. az adatbázis tényleges létrehozása,
iv. az adatbázis online felületen elérhetővé tétele,
v. az adatbázis folyamatos frissítése.

Jelen előadásunkat a ii. kérdésnek szenteljük, amely a nyelvtechnológiai kihívásokat tartalmazza.

Jelenleg egy olyan 199 folyóiratot számoló mintán dolgozunk, amely a Magyarországon kiadott társadalomtudományi folyóiratok átfogó metszetét teszi ki. Az eddig összegyűjtött anyag mintegy 34 ezer tanulmány, ami akár kézi munkával is feldolgozható lenne, ám az adatbázis folyamatos bővítésének igénye elengedhetetlenné teszi a referenciák kinyerésének automatizálását.

A hivatkozások kinyerését két szakaszban végezzük. Az első lépés során a folyóiratszövegekből kivesszük a hivatkozásokat, a második lépés alatt elvégezzük ezek annotációját. Az első lépésben a következő nehézségek merülhetnek fel, amelyek a ráfordított többletmunka miatt jelentős mértékben lassíthatják a feldolgozást (egy nyomós érv a stíluslapok konzekvens betartása mellett): a referenciák nem csak szövegvégi helyzetben fordulnak elő, a tanulmányvégi közlés mellett gyakori a lapalji hivatkozás (technikailag „lábjegyzetelés”), valamint előfordul még szövegközi megjelenítés is (a folyó szövegben zárójelekkel elhatárolva találhatók a hivatkozott publikációra vonatkozó adatok). A hivatkozásokat tehát nem egyszerűen a folyóiratbeli tanulmányvégi pozíciójuk, hanem jellegzetes alkotóelemeik (személynevek, évszámok, tipikus funkciószavak – pl.: in, szerk. – stb.) és szintaktikai felépítésük (az elemek sorrendje, a köztük lévő írásjelek) teszik felismerhetővé.

A második lépés a szövegből kinyert hivatkozások feldolgozása. Ez az első feladatnál jóval bonyolultabb, mivel ekkor már nem szorítkozhatunk csupán az egyes jellemzők felismerésére: itt a teljes hivatkozás pontos elemzésére van szükség (ami technikailag a folyó szövegek XML-annotációját jelenti). Az egyes elemek elemzésekor azonosítani kell egyrészt a hivatkozás típusát (könyv, fejezet, folyóirat stb.), másrészt az adott típushoz tartozó jellegzetes hivatkozáselemeket (pl. név, dátum, cím, kiadás helye, oldalszám stb.). Ez utóbbi nem egyszerű, főleg az egyes elemek közti elválasztóelemek folyamatos variációi, illetve a különféle stíluslapok megléte, valamint az egyéni hivatkozásstílusok használata miatt [5]. Az annotációt első körben

XML-elemekkel, illetve azok konverziója után a BibTex-szabvány címkéivel végezzük (bár némely esetben – pl. a név elem – utóbbinál részletesebbek vagyunk).

3 A megoldás

3.1 A szoftver

A hivatkozások szintaktikai elemzéséhez a szabad felhasználású NooJ szoftvert használjuk, mely egy, a grammatikai elemzés támogatására létrehozott fejlesztőkörnyezet [7]. Munkánk szempontjából a NooJ legfontosabb előnyeit a következőképpen lehet összegezni:

- moduláris felépítése révén alkalmas az egyes elemek lokalizálására,
- az egyes modulok és gráfok könnyedén kombinálhatók és a célnak megfelelően módosíthatók,
- grafikus ábrázolásmódjának köszönhetően átláthatóbb a más szövegkinyerésre és szövegfeldolgozásra alkalmas programoknál,
- gyors,
- kombinálható a NooJ-ban íródott magyar szintaktikai elemzővel,
- az elemzésbe szótárak is bevonhatók (ez fontos lehet például a magyar személy- és keresztnevek, a kiadók esetében vagy például az egyes hivatkozáselemek identifikálásában).

A hivatkozások annotálása közben külön-külön NooJ-grammatikák ismerik fel a jellegzetes hivatkozáselemeket (pl. név, dátum, cím, kiadás helye, oldalszám stb.), majd ezek megfelelő szintaktikai kombinációi illeszkednek a hivatkozások különféle típusaira. Az elemzési szempontokat tartalmazó algoritmusokból, ún. lokális grammatikákból állnak, amelyek a NooJ grafikus felületén szerkeszthetők (l. az 1. és 2. ábrát). A gráfok – elemzésünkben – elsősorban karaktersorokat meghatározó mátrixokat, illetve szótárakat tartalmaznak, de alkalmasak morfológiai és szintaktikai információk tárolására, illetve visszakérésére is. Az elemzés (részleges vagy teljes) találatairól a program konkordancialistát készít, amelyben az egyes találatokat a gráfokban kódolt XML-elemekkel lát el.

3.2 Az eljárás

Az annotálás automatizálása attól válik izgalmas feladattá, hogy az egy típushoz tartozó hivatkozások is roppant sokfélék lehetnek. Eltérő lehet a szövegelemek sorrendje, illetve az ezeket elválasztó írásjelek használata.

A fő kérdés tehát az, hogy melyek azok a legfőbb vonások, amely alapján az emberi intelligencia képes típusokba sorolni és annotálni egy-egy hivatkozást. A felismerőmodulok létrehozásakor abból a feltevésből indulunk ki, hogy a feladatot pusztán felszíni formai mintákra támaszkodva, a jelentésre való bárminemű hivatkozás nélkül kell megoldanunk. A puszta karaktersorozatok felismeréséhez a tesztek

folyamán nem használtunk szótárakat, ugyanakkor az egyes NooJ-modulok összeállításakor bizonyos elemeket előre bekódoltunk (pl. a hivatkozások egyes elemeinek felismerésében segítő *eds., and, &, In:, in:, in* stb. formánsokat). A bibliográfiai tételek egyes elemei (dátum, névkifejezés, bizonyos határoló központozó jegyek) viszonylag jól felismerhetőek, bár a határoló elemek inkonzekvens használata, illetve azok lehetséges kombinációi okoznak némi fejfájást. Mindemellett az is lényeges, hogy a hivatkozáson belül a cím felismerése csakis annak teljes körű szemantikai értelmezése révén lenne elvégezhető, ez pedig meghaladja jelenlegi tudásunkat (bár a többi elem felismerése nagyban segíti a cím felismerését is).

3.3 NooJ-grammatikák

Összhangban a hasonló témájú nemzetközi kutatásokkal [5] – a következő grammatikákat hoztuk létre:

- „név" elem: keresztnév, vezetéknév, eseleges középső név előfordulásának, a köztük levő összekötő elemeknek, valamint a szerzőség típusára (szerző, szerkesztő, szerkesztők) utaló elemek különféle változatai
- évszám: a publikáció kiadásának évére utaló információk (számok és egyéb összekötő karakterek, valamint betűk összessége)
- kiadó neve: a kiadó(k) nevét lefedő karaktersorozatok (betűk és köztük lévő kapcsolatok)
- kiadás helye: a kiadás helyére vonatkozó karakterkombinációk összessége
- oldalszámok: oldalszámok és határoló elemeiknek összessége
- fejezetcím: a fejezetcím elemeit összesítő gráf (a különféle változatok miatt – kisbetű, nagybetű, különféle határoló karakterek, mint a pont, vessző – a cím mező pontos behatárolása szinte lehetetlen)
- könyvcím: a fejezetcímhez hasonló elemek halmaza, amelyek detektálásában a cím melletti elemek segítenek
- évfolyamszámozás: az évfolyam különféle megjelenítésének módozatai

Az egyes grammatikák közül a „cím" mező annotálása a legnehezebb. A benne megjeleníthető bármilyen karakter (betű, szám és határoló írásjelek[2]) miatt szinte lehetetlen pontos grammatikát írni. Mindaddig azonban, amíg pontosan összeállítjuk a mellette álló grammatikák szerkezetét, a „cím" mező is viszonylag sikeresen parszolható: ez azonban inkább a pozícióból, mint a grammatika helyes szerkezetéből adódik.

Különböző problémák forrása lehet a hivatkozás „név" eleme is. Bár a „név" mező után a hivatkozások általában határolóelemmel (vessző, pont, zárójel) vannak elkülönítve, az egyes határolóelemek mégsem határolnak eléggé. Ennek fő oka, hogy az itt megjelenő elemeket több helyen is használják. A helyes parszolás számára támpontot nyújthatna a potenciális mezők egymásutánisága is. A név elem egyértelműsítésében legnagyobb segítség lehetne a „név" után következő „évszám" – azaz „névnek" minő-

[2] Cím esetében a többi mezőtől eltérően bármilyen karakter bármilyen kombinációja elképzelhető.

sül, ami a kiadás éve előtt van –, azonban nem minden hivatkozásstílus helyezi a „név" után az „évszámot" (ugyanis előfordul a hivatkozás végén is).

Egy további probléma a „név" elem belső sokszínűsége, összetettsége. A többszerzős tanulmányok jelölése, a keresztnév iniciáléval történő jelölése – főként több keresztnév esetén – és az esetleges középső név vagy iniciálé csak tovább növeli az elemzések bizonytalanságát. Az elemzésnek ugyanakkor fel kell készülnie a határolóelemek sokféleségére, illetve inkonzekvens használatára is (pl. vessző és pontosvessző keverése, vagy csak vessző használata).

A „név" elem bonyolultsága sokszor túlelemzéseket is eredményez. A „név" elem pontosabb bontása (vezetéknév, középső név, keresztnév), valamint a lehetséges találatok számának növekedése miatt a túlelemzések jelentősen megszaporod(hat)nak (csak reguláris kifejezések használatával, szótárak nélkül nehéz pontos találatot kapni).

Az egyes mezők pontosabb meghatározásában nagy segítség lehet a mezőtípusokra jellemző elemek beazonosítása (az viszont már problémaként merül fel, hogy a *vessző*, *pont* és *pontosvessző* határolóelemként is szerepelhet – ráadásul stílusonként változó szerepben). Különféle hivatkozásstílusok alapján végzett elemzések azt mutatják, hogy a hivatkozások belső struktúrájában összesen 13-féle jelölő használható fel, amelyek detektálása megkönnyítheti a szabályalapú elemzést[3]:

1. táblázat: Jelölők típusai.

Határolóelemek	Mező
1. Vessző ,	név, évfolyam, szám, oldalszám *határolóelem
2. Pont .	név, oldalszám *határolóelem
3. Pontosvessző ;	név *határolóelem
4. Kettőspont :	évfolyam, szám, oldalszám
5. Gondolatjel – -	évfolyam, szám, oldalszám
6. Kötőjel -	évfolyam, szám, oldalszám
7. Kerek zárójel () ( )	év, évfolyam, szám, oldalszám
8. Szögletes zárójel []	sorozat
9. Idézőjel „" " " " "	cím
10. Három pont…	cím
11. Kérdőjel ?	cím
12. Felkiáltójel !	cím
13. Aposztróf '	név, cím

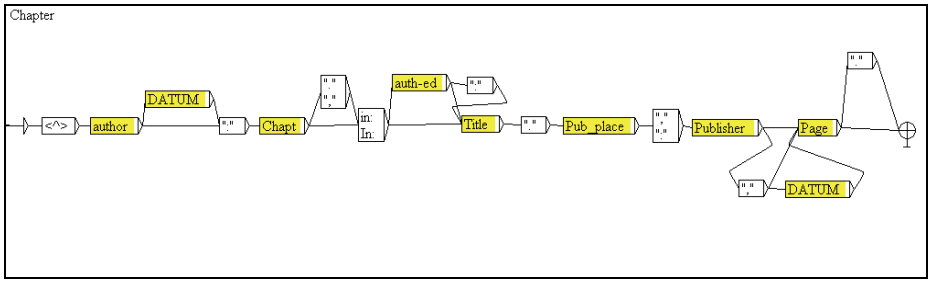
A grammatikák által lefedett mezők bizonyos, a grammatikára/hivatkozáselemre jellemző adatok – sztringsorok – alapján a hivatkozások egyértelműsíthetők. Ezek

[3] Bár a reguláris kifejezések használata elterjedt módszer a hivatkozások annotálására, egyes tapasztalatok szerint bonyolultabb struktúrák parszolására és rendezésére teljes mértékben alkalmatlanok [6].

elsősorban a névre, kiadás évére és helyére, valamint az oldalszámra vonatkozó adatok (pl. *eds*, *vol*, *in*, *kettőspont és szám kombinációja*).

Tesztjeink folyamán az alapelemek összekapcsolásával a következő hivatkozástípusokat parszoltuk teljes vagy részleges lefedettséggel: könyv, fejezet/tanulmány, folyóirat/folyóiratcikk, konferenciakötet – előadásunk a könyv és folyóirat-tanulmány felismerésének részeredményeit mutatja be. A tesztelés során szabadon választott 1027 darabból álló hivatkozásmintán a fenti elemek mellett a következő típusok fordultak elő: előadás, lexikon/szócikk lexikonban, szótár, kézirat/megjelenés alatt, diszszertáció/thesis.

A moduláris felépítés előnye, hogy az egyes grammatikák elemei beágyazhatók más-más grammatikákba (például a „név” mező, amely minden hivatkozástípusban azonos, és minden hivatkozástípusban előfordul), miáltal folyamatos fejlesztésük az összes hivatkozástípusra hatással van. Így a tesztek során, a találati lista lefedettségének növelése céljából azok folyamatosan módosíthatóak (pl. újabb – akár egyedi – hivatkozásstílus megjelenésekor). Természetesen ugyanazon elem többször is beágyazható ugyanazon grammatikába, akár opcionális változóként is (lásd a 1. ábra „dátum” elemét).



1. ábra. A „chapter” (fejezet) hivatkozást parszoló grammatika ábrája: a fő nodsor alatt és fölött lévő nodok kapcsolása fakultatív.

Mivel a hivatkozások egyes elemei formailag megegyeznek, megegyezhetnek (pl. a cím és az alcím, a tanulmány és az azt tartalmazó könyv szerzője), a parszolásban nemcsak a nodok tartalma (azaz a sztringsor és/vagy szótár), hanem az egyes elemek sorrendje is identifikál: pl. az 1. ábra „pubplace” eleme formailag megegyezik a „publisher”, vagy a „chapter” elemével, annotálásnál azonban a sorrend egyértelműsíti azt (az egyes grammatikák összetettsége, a határolóelemek többszöri előfordulása hosszabb – és egyben bonyolultabb – hivatkozásoknál csökkenti a parszolás eredményességét, elsősorban a határolóelemek hivatkozásmezőkben való előfordulása miatt).

Azonos hivatkozástípusok többféle realizációja (többféle stíluslap, illetve egyedi megoldások) miatt és a lehető legjobb eredmény elérése érdekében a fő hivatkozástípusoknak többféle változata is megjeleníthető: ezeket az elemzés XML-kimenete egyértelműsíti[4]. Sajnos a különféle stílusok bizonyos pontjainak átfedése, illetve a hivatkozások szerkezetének viszonylagos strukturálatlansága (a mintegy 200 folyóirat

[4] Az összes grammatika egyszerre futtatható, a keresés eredménye így a (részben vagy teljesen) annotált hivatkozás.

stíluskészlete közel sem nevezhető konzekvensnek – nemhogy egy folyóiraton belül, hanem sokszor egy tanulmányon belül sem) miatt előfordul, hogy a parszolás eredményeképpen ugyanazon hivatkozás több hivatkozástípusba is besorolódik. Ezeket az eseteket jelenleg manuálisan tudjuk csak egyértelműsíteni, azonban a megfelelő szótárak alkalmazása segítség lehet (ez utóbbi összeállítása folyamatban van, elsősorban a kiadók, a folyóiratok, valamint vezeték- és keresztnevek lexikonjainak kiépítését tervezzük). Az egyedi tartalmi és főleg stílusbeli megoldások miatt a grammatikákat folyamatosan építeni és finomítani kell: mindaddig, amíg el nem fogynak a különálló esetek.

4 Az eredmények

A tesztelések előtti kismintás pretesztelések folyamán fejlesztettük ki a jelenleg is használatban lévő grammatikákat (amelyek még tökéletesítésre szorulnak). Tesztjeinket az Akadémiai Kiadó által megjelentetett, általunk szabadon kiválasztott 16 társadalomtudományi folyóirat hivatkozásain végeztük. Az 57 753 db hivatkozásból véletlen mintavétellel kiválasztott 1027 darabos mintán[5] végzett NooJ-tesztelések a következő eredményeket és tapasztalatokat hozták.

Bár a minta nem reprezentatív (kiválasztásában nem játszottak szerepet az alapsokaság – mintegy 33 ezer darab hivatkozás – tulajdonságai), a tudományos folyóiratok leghangsúlyosabb magyarországi kiadójának társadalomtudományi kiadványait tekintve mindenképpen irányadó. A mintának választott hivatkozások típusai kézi szelekció után a következők:

2. táblázat: Teszthivatkozások tipizálása manuális annotálás után.

Hivatkozástípusok	db (N=1027)	% (N=100)
könyv (book)	411	40,0
folyóirat-tanulmány (article)	358	34,9
fejezet vagy tanulmány könyvben (chapter)	164	16,0
egyéb bibliográfiai tétel	94	9,1

A kézi válogatás után az egyes hivatkozáscsoportokon lefuttatott grammatikák a következő eredményt hozták.

Az első tesztelések során azt vizsgáltuk, milyen pontossággal ismerik fel a grammatikák az egyes hivatkozástípusokat. A hivatkozásokon belüli parszolás eredményei a hivatkozás hosszától és annak bonyolultságától függően eléggé szórtak. A hosszabb hivatkozások (elsősorban a már említett „név” és „cím” elemek összetettsége miatt) kevésbé pontosak, gyakoribb az egyes tételek „túlelemzése”.

[5] Az ábécésorrendbe rakott 57 753 hivatkozás minden 50. elemét kiválasztva kapott 1150 tételből manuálisan eltávolítottuk a nem hivatkozási elemeket, azaz a csonka tételeket, vagy a hivatkozások közé került nem referenciákat. Így kaptuk meg az 1027 db hivatkozást.

3. táblázat: Elemzések találatának pontossága.

	∑	T	jó találat (1)	jó találat (2)	pontosság (1) %	fedés (1) %	pontosság (2) %	fedés (2) %
teljes	769	192	87	119	45,31	11,31	61,98	45,31
article	358	33	30	-	90,91	8,38	-	-
book	411	192	89	-	46,35	21,65	-	-

A 3. táblázatban a grammatikák pontosságát és lefedettségét foglaltuk össze. Elemzésünkben csak a könyv és tanulmányok találati pontosságát érintjük, mivel – a már említett okok miatt – egy tételhez gyakran több elemzés is tartozik (192 hivatkozáshoz összesen 692 elemzés, tételenként átlag 3,6). Ezért kérdés, mit tekintünk helyes felismerésnek: ha egy tételnek van legalább egy helyes elemzése, vagy ha minden elemzése helyes (a típus felismerése szempontjából). A „teljes" nyelvtan esetében a tesztmintán egyszerre futtattuk le a rendelkezésünkre álló grammatikákat (az „article" és „book" esetében csak a folyóirat-tanulmányra és könyvre írt grammatikákat teszteltük). Kiértékelésénél szigorúbb módon jártunk el, csak azokat az eseteket fogadtuk el, ahol az adott hivatkozáshoz csak jó elemzések tartoztak. Az egyes grammatikák külön tesztelésénél erre nem voltunk tekintettel. Ezeknél az volt a célunk, hogy az adott ág teljesítményét önmagában nézzük, függetlenül attól, hogy esetleg más grammatikák (más típusba tartozóként) is felismernék az adott hivatkozást.

A „T" oszlop tartalmazza a tesztanyag azon tételeinek számát, amelyekre (legalább egyféleképp) illeszkedik a tesztelt nyelvtan. A „jó találat (1)" oszlop a ∑ oszlop (azaz a kézi annotálás eredményeit tartalmazó oszlop) hivatkozásain lefuttatott grammatikák találatát tartalmazza azon hivatkozásokra, ahol mindegyik elemzés jó volt a kérdéses hivatkozásnál. A „jó találat (2)" oszlop azon elemzések számát tartalmazza, ahol egy tételhez tartozó elemzések legalább egyike jó. A „pontosság (1)" értékei „jó találat (1)" és a „T" százalékos arányát, a „fedés (1)' értékei a „jó találat (1)" és a „∑" százalékos arányát, míg a „pontosság (2)" a „jó találat (2)" és a „T", a „fedés (2)" a „jó találat (2)" és a „∑" százalékos arányát mutatja.

4. táblázat: F-mérték.

	pontosság (1) %	fedés (1) %	pontosság (2) %	fedés (2) %	F-mérték (1) %	F-mérték (2) %
teljes	45,31	11,31	61,98	45,31	18,10	24,76
article	90,91	8,38	-	-	15,35	-
book	46,35	21,65	-	-	29,51	-

A teljes mintán lefuttatott elemzések értékeiből számított kiegyensúlyozás (F-mérték) értéke csupán kis mértékben tér el mindkét elemzés esetében (bár az értékek viszonylag alacsonyak). Abban az esetben, ha azokat a találatokat vesszük figyelembe, ahol az elemzés minden eredménye jó volt (1), a teljes elemzés F-mértéke 18%, míg abban az esetben, ha azokkal a találatokkal foglalkozunk, ahol (a túlelemzés

mellett) legalább egy helyes elemzés született, a teljes elemzés F-mértéke 25%. Az „article" hivatkozásokon lefuttatott, folyóirat-tanulmányok parszolására írt grammatika találati pontossága 15%, míg a „book" hivatkozásokon lefuttatott, könyvek parszolására írt grammatikáé 30%.

Tesztjeink során a NooJ-grammatikák pontossága elmaradt a jóval időigényesebb manuális elemzésnél (ez is erősíti a különféle szótárak összeállítását és az elemzésbe történő bekapcsolását). Az egyes grammatikák sikertelenségébe belejátszott az elemzett hivatkozások sajátosságainak (önálló, a folyóiratra jellemző stylesheet) tükröződése is.

5 Távlatok, tervek, fejlesztések

Bár a kismintás előtesztelések (azaz a grammatikák szerkesztése közbeni próbatesztek) minden esetben sikerekkel zárultak, a viszonylag nagyobb mintán elvégzett tesztek eredményei leginkább a lehetséges korrekciók kidolgozására és újabb módszerek kipróbálására ösztönöznek bennünket. A csupán mintaillesztés alapján működő parszolás úgy tűnik, a magyarországi hivatkozások esetében sem lesz sikeres (a Min-Yuh Day és munkatársai által elvégzett tesztekhez hasonlóan [6]), az elemzésbe szótárakat is be kell fűzni. Szótárakat elsősorban a „név", „kiadó" és „folyóirat" mezők esetében tudunk használni.

Egy további megoldás lehet a „név" mező egyszerűsítése (akár arra hivatkozva, hogy az évszám előtt álló karaktersorozatot tekintjük névnek – mindezt megszorításokkal, mivel nem minden hivatkozás szerepelteti az évszámot a szerző után –, és egy következő elemzés során egyértelműsítjük azt), mivel a tapasztalatok azt mutatják, hogy a „név" pontos parszolására való törekvés jelentős mértékben megnöveli az elemzések számát.

Az eddig elvégzett munkák haszna, hogy előrevetítik a további lépéseket. Amellett, hogy a hivatkozások adatbázisának építése a megfelelő ütemben halad (feldolgoztuk a kiválasztott folyóiratok tartalomjegyzékét és kidolgoztuk az adatbázis struktúráját), tökéletesíteni kell egyrészt a meglévő grammatikákat, illetve ki kell terjeszteni azokat a többi hivatkozástípusra is.

Szakirodalom

1. Bergmark, D.: Automatic extraction of reference linking information from online documents. TR2000-1821 (2000)
2. Constans, P.: Approximate textual retrieval, arXiv:0705.0751 (2007) http://arxiv.org/PS_cache/arxiv/pdf/0705/0705.0751v1.pdf
3. Constans, P.: A Simple Extraction Procedure for Bibliographical Author Field (2009) arXiv:0902.0755. http://arxiv.org/PS_cache/arxiv/pdf/0902/0902.0755v1.pdf
4. Giuffrida, G., Shek, E. C., Yang, J.: Knowledge-based metadata extraction from PostScript files. In: Proceedings of the Fifth ACM International Conference on Digital Libraries (2000) 77–84

5. Day, M.-Y., Tsai, T.-H., Sung, C.-L., Lee, C.-W., Wu, S.-H., Ong, C.S., Hsu, W.-L.: A Knowledge-based Approach to Citation Extraction. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2005). Las Vegas, Nevada, USA. (2005) 50–55
6. Day, M.-Y., Tsai, T.-H., Sung, C.-L., Hsieh, C.-C., Lee, C.-W., Wu, S.-H., Wu, K.-P., Ong, C.S., Hsu, W.-L: Reference metadata extraction using a hierarchical knowledge representation framework. Decision Support Systems Vol. 43 (2007) 152–167
7. Silberztein, M.: NooJ manual. Available at the website http://www.nooj4nlp.net (2003)